Named entity recognition model

by Yuf Azhar

Submission date: 04-Jan-2020 09:58AM (UTC+0700)
Submission ID: 1239204922
File name: Named entity recognition model for Indonesian tweet using CRF classifier (Artikel).pdf (557.56K)
Word count: 2806
Character count: 14846

Named entity recognition model for Indonesian tweet using CRF classifier

Named entity recognition model for Indonesian tweet using CRF classifier

Y Munarko, M S Sutrisno, W A I Mahardika, I Nuryasin and Y Azhar

Teknik Informatika, Universitas Muhammadiyah Malang

yuda@umm.ac.id

Abstract. Named Entity Recognition (NER) is a part of Natural Language Processing (NLP) that acts to recognize the existing word entity in the document. By using NER, it is possible to perform activities such as information extraction and text summary. One of the data sources for the NLP process is tweets which are real time, occurred frequently, but limited by the number of words per tweet. In Indonesia, twitter is one of the most popular social media with various topics, so, it is necessary to provide models, train data, and test data for Indonesian tweet. In this study, the models were built using Conditional Random Field classification from 8,000 tweets that have been grouped to formal tweets and informal tweets. By testing the models to 2,000 training data, it provided recall and precision results of 62% and 87% respectively for formal tweets, 36% and 90% respectively for informal tweets, and 60% and 86% respectively for mixed tweets. These results indicate that the created Indonesian tweet models can be used for automatic NER.

1. Introduction

Twitter is one of the most popular social media in Indonesia with a large number of users. As a micro-blogger, Twitter provides users with the flexibility to tweet a maximum length of 140 characters. This is quite interesting because users tend to share information frequently with a typical sentence structure which is different with common text documents. Information on Twitter is very useful when extracted appropriately, for example, the information extraction on earthquake detection and sentiment analysis in real time [1-3]. For information extraction needs, Named Entity Recognition (NER) plays a very important role. Early NER research generally used documents in English corpus or traditional text documents, but later research concerned on various corpora, such as non-English language, tweets, user review, and biomedical domain [4-10]. Algorithms used for NER are varied but can be classified into two types, those are machine learning based and rule-based. The example of machine learning based is the use of Conditional Random Fields (CRF) [4], then the example of rule-based is MUSE that utilized GATE [11].

In general, machine learning based such as CRF provides good performance almost for all types of corpora. CRF is based on sequence probability that used to build a model. The model, then, is utilized to detect entities automatically. Different with machine learning based, rule-based is a kind of tailor-made solution which usually only work for a particular corpus. The rule is built by analyzing the nature of the corpus and then employs various features such as part of speech, context, and the lexical word [12].

For the Indonesian language, there were several studies on NER, including InNer [12], the use of association rule [13, 14], and the use of supervised learning [15]. Those studies already shown a promising results with high precision and recall values. However, the data used were only standard documents which were using standard sentence structure and standard words. So, there is no evident that those studies may works well on Twitter which uses non-standard sentence structure and non-standard word. Moreover, the rule-based approach may experience suffer performance when apply for a different corpus domain.

Study about NER for microblog already conduct by several scholars, such as Twiner [8] and supervised learning approach [16]. Those studies were concerning on a domain specific corpus. Twinner target was a stream of tweets in English which firstly tokenized into phrases and then each phrase classified to a suitable entity. The second study was tried to detect words' entity for the tweet about labor strike event in Indonesia.

In this research, we built a NER model that can be used to recognize entities in Indonesia tweets automatically using various corpus, so it can be used as a general model for any kind of corpus. We were also utilizing CRF as a machine learning based solution, which suitable for the non-domain specific corpus.

2. Research method

The research was conducted by several steps, begins by data collection, and then follows by data classification, data preprocessing, entity tagging, model development, and testing.

2.1. Data collection

Data collection was conducted over two months from several popular general users, news agencies, and government agencies. The purpose of choosing a tweet based on the type of user is to get a collection of formal and informal tweets. Popular general users usually make informal tweets, while news agencies and government agencies tend to make tweets with more formal grammar and formal word selection. Tweets collected as many as 10,000 tweets that will be divided into train data and test data, and formal tweet and informal tweet.

2.2. Data classification

For testing purposes, the tweet was divided into two datasets, i.e. train data, and test data. Train data is used to build a model, while test data is used to perform model testing. The two datasets are then grouped by linguists into two groups. The first group contains tweets that use formal Indonesian, while the second group contains tweets that use informal Indonesian. The classification of data is described in table 1.

Table 1. The data used.

Type of data	Indonesian Tweets		#tweets
	Formal	Informal	
Train data	3,814	4,186	8,000
Test data	1,000	1,000	2,000
Total			10.000

2.3. Data processing

Before tagging entities manually, tweets are first preprocessed. Standard preprocessing is to convert characters into lower case and tokenizing. However, these simple preprocessing processes become quite complicated on twitter. For example, the tokenizing process cannot be based on white space only, because there are some words that are not separated by whitespace. To simplify, in this study, the tokenizing process was performed manually. Furthermore, the characters in a tweet are not just letters, such as emoticons, so, non-letter characters were left as is. The entity tagset is shown in table 2.

Table 2. Entity tagset.

#	Tag	Description	Example
1	Person	Name of person	Ahok, Ridwan Kamil
2	Location	Name of location	Jakarta, rumah, teras
3	Organization	Name of organization	BMKG, BNI, BNN
4	O	Other	yang, di, ke

2.4. Entity tagging

Tagging entities for train data and test data were done manually against 10,000 tweets. As can be seen in table 2, there are four entity tag, Person, Location, Organization, and O. All tokens are classified into one of the four entities in table 2. Examples of manual tagging results are presented in table 3.

Table 3. Tweet examples.

#	Tweets	Tweet type	Tagged tweets
1	Masihkah Madrid Merasa Kehilangan Di Maria ?	Formal	Masihkah/O Madrid/ORGANIZATION Merasa/O Kehilangan/O Di/PERSON Maria/PERSON ?/O
2	Densus 88 Tangkap Warga Klaten , Sita Motor dan Buku Bertema Radikalisme https://t.co/lnY8xs6VRF	Formal	Densus/ORGANIZATION 88/ORGANIZATION Tangkap/O Warga/O Klaten/LOCATION ,/O Sita/O Motor/O dan/O Buku/O Bertema/O Radikalisme/O https://t.co/lnY8xs6VRF/O
3	Nyeblak (Evisahara, Avis Yuni)	Informal	Nyeblak/O (/O Evisahara/PERSON ,/O Avis/PERSON Yuni/PERSON)/O
4	3 in 1 bakal dihapus brooo https://t.co/eIsdguQnXO	Informal	3/O in/O 1/O bakal/O dihapus/O brooo/PERSON https://t.co/eIsdguQnXO/O

2.5. Model development

The tagged train data was then processed using CRF classifier, resulting three models. In order to generate these model, we were utilizing Stanford NER which already implemented CRF classifier [4]. The first model is based on format tweets, the second model is based on informal tweets, and the last one is the combination of both formal and informal tweets.

2.6. Testing

The test was performed to measure the recall and the precision values. The recall is the comparison between the token that correctly identified, I_c, and the number of tokens that should be correctly identified, I_r, shown as equation (1). Then, the precision is the comparison between the token that correctly identified, I_c, and the number of tokens that identified as correct, I_{db}, shown as equation (2).

$$Precision = \frac{I_c}{I_r} \quad (1)$$

$$Precision = \frac{I_c}{I_{db}} \quad (2)$$

3. Result and discussion

Each model was tested for three different test data, formal, informal and mixed. The purpose of these tests was to see the significance of classifying tweets to formal or informal classes. In addition, precision and recall measurements were performed with ten-fold cross-validation to determine the performance of the model when it does not consider formal and informal classification. The results of the tests are shown in table 4.

We found that the precision value for the test of all test data using all models is of high value. However, selecting the right type of model against test data will result in better performance. For example, the highest precision value for formal test data was obtained by using a formal model, as well as for informal test data and mixed test data. For those tests, the precision values are 0.8760, 0.9076 and 0.8749 consecutively. This indicates that the use of CRF for NER detection on Twitter in Indonesian is of high accuracy.

Moreover, it seems that the completeness of detection was moderately for formal and mixed test data, but slightly low for informal data. This indicated by the highest recall values which were 0.6229, 0.6221 and 0.3601 for formal, mixed, and informal test data respectively. The problem with twitter data is the vast use of non-standard words, both informal tweets and even more in informal tweets. Thus, it is recommended to normalize the word on Twitter using a dictionary or certain rules, so that detection completeness increases. When compared to the results of the entity detection of Indonesian document in other studies, there is an indication that the use of CRF has a better performance. For example, a study that incorporated context, part of speech, and lexical features to formulate rules for detecting entities in formal Indonesian documents, the recall was 0.6343 and the precision was 0.7184 [12].

Table 4. Test result

Model	Test data	Recall	Precision
Formal	Formal	0.6229	0.8760
	Informal	0.3922	0.8121
	Mixed	0.5400	0.8699
Informal	Formal	0.2866	0.8272
	Informal	0.3601	0.9076
	Mixed	0.3339	0.8749
Mixed	Formal	0.6221	0.8749
	Informal	0.5576	0.8523
	Mixed	0.6080	0.8684
Mixed, 10-fold cross validation	Mixed	0.6906	0.8774

In general, we may agree that in order to detect entity correctly, firstly a tweet should be classified into a formal or informal tweet. Secondly, the entities in the tweet can be detected using a suitable model. In a case that tweet classification is not performed, the mixed model may perform very well since the precision and recall values of the test of the mixed model with the mixed train data are only slightly under the precision and recall values of the test of the formal model with the formal train data. To make sure this condition, we measured precision and recall of mixed data using ten-fold cross-validation, which the results are 0.8774 and 0.6906 respectively, which is higher than other measurements.

4. Conclusion

Based on test results, CRF classifier can be used for NER with good performance, where the recall and precision values are respectively 62% and 87% for formal tweets, 36% and 90% for informal tweets, and 60% and 86% for mixed tweets. Thus, the selection of the model has an effect on automatic recognition accuracy, therefore, it is necessary to classify tweets into a formal or informal then

selected the appropriate model for NER. Under conditions where it is not possible to classify tweets, the use of mixed models is adequate, evidenced by 69% of recall value and 87% precision of value in 10-fold cross-validation test.

References

[1] Sakaki T, Okazaki M and Matsuo Y 2010 Earthquake shakes twitter users: real-time event detection by social sensors *Proceedings of the 19th international conference on World wide web. ACM* pp 851–60

[2] Earle P S, Bowden D C and Guy M 2012 Twitter earthquake detection: earthquake monitoring in a social world *Annals of Geophysics* vol 54 no 6

[3] Wang H, Can D, Kazemzadeh A, Bar F and Narayanan S 2012 A system for real-time twitter sentiment analysis of 2012 US presidential election cycle *Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics* pp 115–20

[4] Finkel J R, Grenager T and Manning C 2005 Incorporating non-local information into information extraction systems by gibbs sampling *Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics* pp 363–70

[5] Leonandya R A, Distiawan B and Praptono N H 2015 A semi-supervised algorithm for indonesian named entity recognition *Computational and Business Intelligence (ISCBI), 2015 3rd International Symposium on. IEEE* pp 45–50

[6] Alfina I, Manurung R and Fanany M I 2016 Dbpedia entities expansion in automatically building dataset for indonesian ner *Advanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on. IEEE* pp 335–40

[7] Suwarningsih W, Supriana I and Purwarianti A 2014 Imner indonesian medical named entity recognition *Technology, Informatics, Management, Engineering, and Environment (TIME-E), 2014 2nd International Conference on. IEEE* pp 184–8

[8] Li C, Weng J, He Q, Yao Y, Datta A, Sun A and Lee B-S 2012 Twiner: named entity recognition in targeted twitter stream *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012* pp 721–30

[9] Kazama J, Makino T, Ohta Y and Tsujii J 2002 Tuning support vector machines for biomedical named entity recognition *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3. Association for Computational Linguistics* pp 1–8

[10] GuoDong Z and Jian S 2004 Exploring deep knowledge resources in biomedical name recognition *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. Association for Computational Linguistics* pp 96–9

[11] Maynard D, Tablan V, Ursu C, Cunningham H and Wilks Y 2001 Named entity recognition from diverse text types *Recent Advances in Natural Language Processing 2001 Conference* pp 257–74

[12] Budi I, Bressan S, Wahyudi G, Hasibuan Z and Nazief B 2005 Named entity recognition for the indonesian language: combining contextual, morphological and part-of-speech features into a knowledge engineering approach *Discovery Science* Springer pp 57–69

[13] Budi I and Bressan S 2003 Association rules mining for name entity recognition *Web Information Systems Engineering, 2003. WISE 2003. Proceedings of the Fourth International Conference on. IEEE, 2003* pp 325–8

[14] Budi I and Bressan S 2007 Application of association rules mining to named entity recognition and co-reference resolution for the indonesian language *International Journal of BusinessIntelligence and Data Mining* vol 2 no 4 pp 426–46

[15] Wibawa A S and Purwarianti A 2016 Indonesian named-entity recognition for 15 classes using ensemble supervised learning *Procedia Computer Science* vol 81 pp 221–8

[16] Purwarianti A, Madlberger L and Ibrahim M 2016 Supervised entity tagger for indonesian labor strike tweets using oversampling technique and low resource features *TELKOMNIKA (Telecommunication Computing Electronics and Control)* vol 14 no 4 pp 1462–71

Named entity recognition model

ORIGINALITY REPORT

10%	8%	12%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	mafiadoc.com Internet Source	5%
2	Deni Cahya Wintaka, Moch Arif Bijaksana, Ibnu Asror. "Named-Entity Recognition on Indonesian Tweets using Bidirectional LSTM-CRF", Procedia Computer Science, 2019 Publication	3%
3	Submitted to University of Edinburgh Student Paper	3%

Exclude quotes	Off	Exclude matches	< 3%
Exclude bibliography	Off		